

Calibra: Dataset Observability and Coreset Selection for Robot Imitation Learning

Technical Report

Ömer Topçu
github.com/omerTT/Calibra

June 2026

Abstract

Robot learning labs routinely train policies on thousands of demonstration episodes without inspecting the data. We present **Calibra**, an open-source dataset observability toolkit that applies four deterministic mathematical estimators — jerk spike rate, velocity discontinuity rate, timing jitter, and log dimensionless jerk (LDLJ) — to detect quality anomalies before training begins. We profile 15 public imitation-learning datasets and find three actionable findings: (1) scripted motion planners produce 22–25% jerk spike rates, 25× higher than human teleoperation counterparts; (2) diverse hardware fleets (DROID) exhibit 10× more velocity discontinuities than single-platform datasets (ALOHA); and (3) control frequency is a latent confounder that inflates velocity discontinuity by up to 80 percentage points in coarse-grained datasets (BridgeData V2, 5 Hz). Calibra is pip-installable and runs on any dataset with a LeRobot v2 adapter in under two minutes.

1 Introduction

Imitation learning pipelines treat data collection as a solved problem. Demonstrations are recorded, concatenated into a dataset, and fed to a training script. Quality inspection — if it happens at all — is manual, inconsistent, and expensive.

In practice, real demonstration datasets contain three classes of silent problems. **Mechanical noise:** jerk spikes from communication lag, stiff joints, or abrupt operator corrections. **Temporal artifacts:** dropped frames and duplicate timestamps from simulation engines under load. **Distributional skew:** 60–80% near-duplicate episodes in large collections, inflating compute cost and reinforcing dominant modes.

None of these are detectable by standard loss

curves or policy rollouts during training. By the time a trained policy stalls or flails, the practitioner has no way to determine whether the fault lies in the architecture, the training recipe, or the data.

Calibra addresses the data side. It computes *per-episode* and *per-dataset* diagnostic statistics, compares them against a library of validated reference profiles, and flags anomalies with bootstrap confidence intervals. Its four estimators are closed-form, require no learned model, and run in $O(N)$ time over the episode length.

Contributions.

1. Four reproducible data-quality metrics with falsifiable claim thresholds validated across 15 public datasets.
2. An empirical cross-dataset study revealing control-frequency confounding and hardware-diversity effects.
3. An open-source toolkit (MIT licence) with a LeRobot v2 adapter, a `compare` command, and a coreset pruner.

2 Method

2.1 Jerk Spike Rate

For an episode of T steps with action sequence $\mathbf{a}_1, \dots, \mathbf{a}_T$, the per-step jerk magnitude is

$$j_t = \|\mathbf{a}_{t+1} - 2\mathbf{a}_t + \mathbf{a}_{t-1}\|_2.$$

A step is classified as a *spike* if $j_t > 5 \cdot \text{median}(j)$. The episode-level spike rate is the fraction of steps classified as spikes. This threshold is calibrated so that smooth 50 Hz ALOHA teleoperation produces rates below 1% and abrupt planner waypoints exceed 20%.

2.2 Velocity Discontinuity Rate

$$\text{vd}_t = \|\mathbf{a}_{t+1} - \mathbf{a}_t\|_2 / \max_s \|\mathbf{a}_s\|_2$$

Steps where $\text{vd}_t > 0.20$ count as discontinuities. The 20% threshold was chosen to distinguish genuine control-loop instabilities from normal deceleration.

2.3 Timing Jitter (CV)

Given inter-step timestamps $\Delta t_1, \dots, \Delta t_{T-1}$, the jitter coefficient of variation is $\sigma_{\Delta t} / \mu_{\Delta t}$. Values above 10^{-2} indicate clock synchronisation problems; values above 0.1 indicate severe frame drops.

2.4 Log Dimensionless Jerk (LDLJ)

LDLJ [1] is a scale-invariant smoothness index for movement primitives:

$$\text{LDLJ} = \log \left(\frac{D^3}{v_{\max}^2 T^5} \int_0^T \|\ddot{\mathbf{a}}\|^2 dt \right)$$

where D is movement amplitude and $v_{\max}$ is peak speed. More-negative values indicate less smooth motion. LDLJ is computed numerically from the discrete action sequence via finite differences.

3 Experimental Setup

We profile 15 publicly available datasets, all accessible via the LeRobot v2 API. Datasets span simulation and real hardware, human teleoperation and scripted planners, and control frequencies from 5 Hz (BridgeData V2) to 50 Hz (ALOHA). Table 1 summarises the collection.

Table 1: Datasets profiled in this study.

Dataset family	Mode	N	Hz
ALOHA sim human ($\times 2$)	pos	50	50
ALOHA sim scripted ($\times 2$)	pos	50	50
ALOHA static ($\times 4$)	pos	49–50	50
ALOHA mobile ($\times 2$)	pos	18–85	50
DROID-100	pos	100	15
BridgeData V2	vel	50 415	5
PushT / PushT-image	vel	206	10
SVLA SO-100 ($\times 2$)	pos	50–56	50

All metric computations are deterministic and single-threaded; a 100-episode dataset at 50 Hz takes under 15 seconds on a 2022 MacBook Pro.

4 Results

4.1 Finding 1: Scripted planners create artifact noise ($25\times$)

ALOHA simulation offers a controlled comparison: the *same* bimanual task executed by both human operators and a scripted planner at the same frequency and action dimensionality. Human-collected episodes show a mean jerk spike rate of 0.7–0.9%; scripted episodes show 21.9–24.9%. The $25\times$ gap cannot be attributed to task difficulty — it reflects the piecewise-linear trajectory profiles produced by the planner’s waypoint interpolation. Policies trained on scripted data may learn to reproduce planner artifacts that do not appear in real deployment.

4.2 Finding 2: Hardware diversity amplifies noise ($10\times$)

DROID-100 collects 100 episodes across diverse hardware platforms, operator backgrounds, and tasks. It shares the same control mode (position command) as ALOHA hardware, but its mean jerk spike rate is 4.55% and velocity discontinuity is 7.08% — roughly $10\times$ the ALOHA hardware baselines (0.42% and 0.84%). These are not pathological datasets; they represent the realistic operating range of multi-robot, multi-operator data collection. Calibra’s `compare` command surfaces this gap automatically and flags the 13 DROID episodes whose spike rate exceeds 8% as outliers.

4.3 Finding 3: Control frequency is a latent confounder

BridgeData V2 contains 50 415 real-hardware demonstrations at 5 Hz. Its velocity discontinuity rate is 80.5% — far above any other position- or velocity-command dataset we profiled. This is not hardware noise: at 5 Hz, any smooth continuous trajectory appears as a sequence of abrupt step changes when discretised, because the time resolution is too coarse to capture within-step variation. This finding formally falsified claim VD-002 (“velocity-command datasets: 10–20%”) and prompted the frequency-qualified successor VD-003.

The lesson is methodological: *velocity discontinuity rate is not comparable across datasets without controlling for control frequency*. Calibra’s reference profiles record the control frequency alongside each metric, and the `compare` command warns when the target dataset’s frequency differs by more than a factor of two from the reference.

4.4 Finding 4: Timing quality is uniformly excellent

Jitter CV is below 10^{-4} across all 15 datasets, and timestamp dropout is zero in every profile. This suggests that LeRobot v2’s normalised Parquet format provides reliable temporal alignment, and that clock synchronisation issues documented in older Isaac Lab workflows are absent in the published HuggingFace datasets.

4.5 Quantitative summary

Table 2 reports mean spike rate, velocity discontinuity, jitter CV, and LDLJ for all 15 datasets.

5 Claim Registry

Calibra maintains a falsifiable claim registry in `calibra/claims/`. Each claim is a JSON record specifying the metric, the dataset class, the assertion, a formal falsification condition, and an evidence list. The registry currently contains 12 active claims across four metric families, backed by 15 reference profiles (ratio ≥ 1 , per project policy).

Two claims have been formally falsified by the experiments in this paper (VD-002, JS-002), demonstrating that the registry is live empirical science rather than documentation. The falsification events produced two successor claims (VD-003, JS-003) with refined scope.

6 Related Work

Dataset quality for robot learning has received growing attention. Walke et al. [2] document BridgeData V2 but do not report control-smoothness metrics. Khazatsky et al. [3] profile DROID for diversity but not for temporal or kinematic quality. Calibra provides the first systematic cross-dataset comparison on the metrics that directly affect policy learning stability.

Data-curation tools such as GRAPE [4] and SOAR [5] focus on reward-based episode selection; Calibra is complementary, operating entirely on raw kinematic data without requiring a reward model or environment access.

7 Conclusion

We have shown that three systematic quality problems — planner artifacts, hardware-diversity noise, and control-frequency confounding — are detectable

from raw kinematic data in under two minutes, before any policy training takes place. Calibra’s deterministic estimators and falsifiable claim registry provide an auditable, reproducible quality baseline for any dataset with LeRobot v2 support.

Future work includes: (1) extending the adapter library to Isaac Lab and ROS 2 bag formats; (2) integrating the coreset pruner with the observability metrics to produce quality-weighted coresets; (3) running a controlled ablation study measuring downstream policy performance as a function of Calibra’s quality scores.

Acknowledgements

This report was prepared with AI writing assistance (GitHub Copilot). All experimental results, metric computations, and dataset profiles are the sole work of the author.

References

- [1] S. Balasubramanian, A. Melendez-Calderon, and E. Burdet. A robust and sensitive metric for quantifying movement smoothness. *IEEE Trans. Biomed. Eng.*, 59(8):2126–2136, 2012.
- [2] H. Walke et al. BridgeData V2: A dataset for robot learning at scale. In *CoRL*, 2023.
- [3] A. Khazatsky et al. DROID: A large-scale in-the-wild robot manipulation dataset. *arXiv:2403.12945*, 2024.
- [4] J. Hejna and D. Sadigh. Few-shot preference learning for human-in-the-loop RL. In *CoRL*, 2023.
- [5] Y. J. Ma et al. Eureka: Human-level reward design via coding large language models. *arXiv:2310.12931*, 2023.

Table 2: Calibra observability metrics across 15 public imitation-learning datasets. Scripted datasets are marked †. Spike and VelDisc are mean per-episode rates (%). LDLJ is mean log dimensionless jerk (more negative = less smooth).

Dataset	Ctrl	N	Spike %	VelDisc %	Jitter CV	LDLJ
aloha_mobile_cabinet	position	85	0.42	0.84	3.0e-05	-24.70
aloha_mobile_shrimp	position	18	0.46	0.66	8.5e-05	-27.80
aloha_sim_insertion_human	position	50	0.69	2.39	1.1e-05	-20.43
aloha_sim_insertion_scripted†	position	50	24.90	0.75	4.0e-06	-17.19
aloha_sim_transfer_cube_human	position	50	0.94	4.56	4.0e-06	-20.02
aloha_sim_transfer_cube_scripted†	position	50	21.87	0.75	4.0e-06	-17.23
aloha_static_battery	position	49	0.44	5.46	1.4e-05	-22.10
aloha_static_candy	position	50	0.42	2.66	1.6e-05	-22.28
aloha_static_coffee	position	50	0.17	1.95	2.6e-05	-24.14
aloha_static_cups_open	position	50	0.30	6.44	4.0e-06	-20.51
BridgeData_V2	velocity	50,415	0.71	80.48	1.0e-06	-13.79
droid_100	position	100	4.55	7.08	6.0e-06	-19.39
pusht_image	velocity	206	7.76	11.11	3.0e-06	-16.01
pusht	velocity	206	4.94	16.72	3.0e-06	-16.34
svla_so100_pickplace	position	50	1.09	6.73	9.0e-06	-20.11
svla_so100_stacking	position	56	1.02	1.65	9.0e-06	-19.61